RegionMoFE: Task-oriented Urban Region Representation Learning with Mixture of Fusion Experts

Anonymous Author(s)

Abstract

The proliferation of multi-source urban data from Web platforms offers unprecedented opportunities for urban computing. However, learning universal and semantically rich region representations that capture complex urban dynamics remains a critical challenge. Early approaches aimed to learn a single, general-purpose representation through one-stage fusion. More recently, to enhance task-specific adaptability, two-stage methods that decouple fusion from adaptation have gained prominence. However, these prevailing methods face two critical, concurrent limitations: the representation learning stage still operates as an uninterpretable "black box," and the subsequent adaptation stage lacks the semantic guidance to align with diverse task requirements. This dual failure results in representations that are both semantically opaque and poorly adapted to downstream tasks. To resolve this, we propose Task-oriented Urban Region Representation Learning with a Mixture of Fusion Experts (RegionMoFE), a principled two-stage framework designed to address both limitations directly. The first stage, Interpretable Universal Representation Learning, achieves interpretable multi-view fusion through a Mixture of Fusion Experts (MoFE) module that operationalizes Partial Information Decomposition (PID) to disentangle heterogeneous interactions into unique, redundant, and synergistic components, yielding structured and interpretable universal embeddings. The second stage, Task-oriented Prompting for Region Embedding (Prompt4RE), builds upon this foundation by leveraging a frozen Multimodal Large Language Model (MLLM) to generate semantically grounded prompts from natural-language task descriptions, enabling guided and explainable task adaptation. Extensive experiments on four diverse tasks demonstrate that RegionMoFE consistently outperforms state-of-the-art baselines. Averaged across all evaluation metrics (MSE, RMSE, and R^2), it achieves improvements of 8.5% in New York City and 15.4% in Chicago, respectively. Our code and data will be made publicly available upon acceptance.

CCS Concepts

• **Information systems → Data mining**.

Keywords

Urban Region Representation, Multimodal Large Language Model, Mixture of Experts

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Anonymous Author(s). 2018. RegionMoFE: Task-oriented Urban Region Representation Learning with Mixture of Fusion Experts. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid proliferation of Web-scale urban data — including user-generated Points-of-Interest (POIs), open mapping platforms, and sensor data — has created unprecedented opportunities for advancing urban computing within the Web ecosystem [18, 24]. This heterogeneous and multi-source data originating from Web platforms and urban sensors poses unique challenges for integration and representation. A central challenge lies in learning unified, semantically rich representations of urban regions that capture both structural and functional characteristics spanning the Web and the physical world. Urban region representation learning thus plays a key role in Web-based semantics and knowledge discovery, enabling downstream tasks such as crime prediction [26] and population forecasting [6]. Ultimately, these representations contribute to building a more intelligent and interconnected Web of cities, empowering data-driven decisions for sustainable and equitable urban development.

Given the inherently multi-source nature of urban data, multi-view representation learning has emerged as the dominant technical approach to address this challenge. A review of these efforts reveals two dominant approaches, each with inherent limitations that motivate our work. The first approach consists of single-stage methods [3, 9, 15, 31], which learn embeddings by fusing multi-view features, often through end-to-end self-supervised training. A primary limitation of these methods is that their fusion mechanisms, such as attention, operate as opaque mechanisms, lacking interpretability. It remains unclear how different data sources contribute to the final representation. This can lead to suboptimal performance on specific downstream tasks that may rely on nuanced, task-specific feature combinations. The second approach includes two-stage, prompt-based methods [14, 32]. These methods first pre-train a general region representation and then introduce lightweight, task-specific prompts for adaptation. The principal challenge here lies in prompt initialization. The prompts are typically initialized randomly or in a task-agnostic manner, meaning the adaptation process lacks effective semantic guidance from the outset. This can hinder the model's ability to effectively tailor the general representation to the unique demands of a specific task.

The limitations of these two approaches can be attributed to a shared underlying limitation: a failure to explicitly model the relationship between multi-view data fusion and task-specific adaptation. The single-stage methods struggle with interpretable fusion, while the two-stage methods lack guided adaptation. This motivates us to address three fundamental research questions:

(Q1) Quantifying Interaction: Can we move beyond treating fusion as a black box and quantitatively measure the distinct types of interaction—namely uniqueness, redundancy, and synergy—that exist among heterogeneous urban data sources?

(Q2) Explaining Decisions: How do these different interaction patterns contribute to model predictions for diverse downstream tasks? For instance, is a high crime rate prediction driven by the unique information from one data view, or a synergistic effect emerging from several?

(Q3) Guided Adaptation: Armed with an understanding of these interactions, can we develop a principled mechanism for task adaptation that is actively guided by task-specific semantics, rather than relying on task-agnostic or random initialization?

To systematically address these questions, we propose Task-oriented Urban Region Representation Learning with a Mixture of Fusion Experts (RegionMoFE), a novel two-stage framework. Our approach is designed to bridge the gap between interpretable data fusion and guided task adaptation.

The first stage, termed Interpretable Universal Representation Learning, focuses on achieving an interpretable, white-box fusion, thereby addressing Q1 and Q2. This stage first generates view-specific embeddings from raw data, which then serve as input to our core innovation, the Mixture of Fusion Experts (MoFE) module. Instead of a single fusion model, MoFE employs a panel of specialized experts. Each expert is trained via weakly-supervised interaction losses to capture a distinct component of information interplay as defined by Partial Information Decomposition (PID): view-specific uniqueness, shared redundancy, and emergent synergy. Governed by an adaptive router, MoFE produces a structured, disentangled universal representation that not only encodes urban features but also reveals their composition, providing a clear window into the model's decision-making process.

The second stage, building on this interpretable foundation, introduces Task-oriented Prompting for Region Embedding (Prompt4RE) to enable semantically-guided adaptation and address Q3. Prompt4RE is a general-purpose module that leverages a frozen Multimodal Large Language Model (MLLM) to generate initial prompt vectors. Guided by a natural language task description, the MLLM extracts high-level semantic concepts relevant to the task. These concepts provide a strong semantic starting point for the prompts. Through a novel Prompt-Representation Alignment (P-R Alignment) mechanism, these initialized prompts then query and activate the most relevant components within the universal and disentangled representation from Stage 1, tailoring the final task-specific embedding for downstream prediction in a principled and interpretable manner.

In summary, our main contributions are: **(1)** We propose RegionMoFE, a principled two-stage framework that unifies interpretable fusion and semantically guided adaptation for urban region representation. **(2)** We design the Mixture of Fusion Experts (MoFE), which operationalizes Partial Information Decomposition (PID) to explicitly disentangle multi-view interactions into unique, redundant, and synergistic components. **(3)** We introduce Prompt4RE, which leverages MLLMs to turn natural-language task descriptions into semantically grounded prompts, enabling guided and interpretable task adaptation. **(4)** We conduct extensive experiments across four diverse downstream tasks, demonstrating significant improvement over state-of-the-art baselines and validating our theoretical principles in practice.

2 Related Work

2.1 Urban Region Representation Learning

Early region representation methods mainly relied on single-modal data [8, 17, 28], yielding limited understanding of the multiply dimensional nature of urban environments. To capture richer interactions across heterogeneous sources, research has moved toward multi-view representation learning, which is broadly categorized into two main paradigms: single-stage and two-stage approaches.

Single-Stage Fusion Methods. These methods create a universal representation via end-to-end fusion, but operate as opaque "black-boxes" that obscure the distinct contributions of each data source. For instance, the dynamic weights in attention-based models (e.g., MVURE [30], HAFusion [15]) do not distinguish between redundant and synergistic interactions. Similarly, graph-based methods like RegionEncoder [7] employ a non-decomposable fusion process. Other strategies, such as contrastive learning (e.g., ReMVC [29]), may even inadvertently suppress valuable view-specific information. In contrast to these approaches, our MoFE module is designed for interpretability, explicitly disentangling multi-view interactions into their unique, redundant, and synergistic components.

Two-Stage Fusion Methods. A more recent paradigm adopts a two-stage approach: (1) pre-training a universal embedding, followed by (2) task-specific adaptation with lightweight prompts. For example, HREP [32] pioneered this direction by using prefix-tuning to adapt embeddings learned from a heterogeneous graph. However, its prompts are learnable but randomly initialized, lacking any initial task-specific semantics. FlexiReg [14] proposes a more advanced two-stage framework, but its "PromptEnhancer" module integrates additional features in a task-agnostic manner. The enhancement is static and not actively guided by the unique semantic demands of a specific downstream task (e.g., "crime prediction" vs. "population forecasting"). To inject richer semantics, another line of work, such as UrbanCLIP [27], leverages MLLMs to generate a generic textual description for each satellite image. However, a common limitation of these methods is that their prompts or semantic enhancements are static and task-agnostic, failing to dynamically align the representation with the specific intent of a downstream task. In contrast, our method is the first to leverage the natural-language description of the task itself (e.g., "predicting crime rates") to generate task-aware prompts, ensuring precise alignment with the downstream objective.

2.2 Information-Theoretic Disentanglement

In the broader field of representation learning, there is a growing movement towards building interpretable models. Techniques rooted in information theory, particularly PID [21], offer a principled framework to decompose the information that multiple variables provide about a target. PID breaks down this information into unique, redundant, and synergistic components. Inspired by these principles, which have seen recent application in explaining neural network behavior [16, 20, 25]. For instance, PIDF [19] leverages PID for feature selection, revealing how each feature relates to the target, while certain approaches use PID to quantify inter-modality

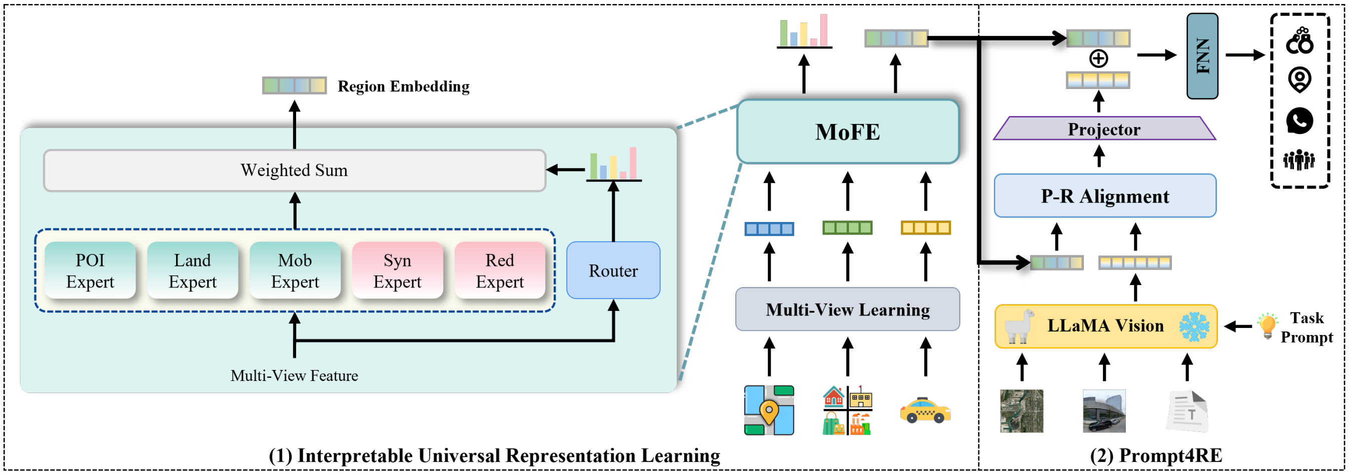

Figure 1: Overall framework of the proposed RegionMoFE.

interactions [10, 25]. Moreover, the PID framework is also utilized to track the redundant and unique information encoded by the hidden neurons with respect to the target variable [16]. However, these approaches cannot be directly applied to explain multi-view fusion in region representation learning, as urban environments exhibit complex spatial dependencies that demand domain-specific architectural designs capable of capturing spatial autocorrelation while preserving interpretability. Distinct from prior works that use PID for analysis or feature selection, our MoFE module is the first to propose a concrete neural architecture that operationalizes PID as a core mechanism for interpretable multi-view fusion, specifically for the domain of region representation.

3 METHODOLOGY

As illustrated in Figure 1, RegionMoFE is a two-stage framework. The first stage, interpretable universal representation learning, aims to overcome the opacity of conventional fusion mechanisms by introducing the MoFE, which explicitly disentangles the unique, redundant, and synergistic interactions among heterogeneous data views. Building upon this interpretable foundation, the second stage, Prompt4RE, enables guided and task-aware adaptation by leveraging a frozen MLLM to generate semantically grounded prompts aligned with natural-language task descriptions. We now detail the architecture and optimization of each stage.

3.1 Preliminaries and Problem Formulation

Let $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ be a set of N disjoint urban regions that partition a city, derived from a standard method such as a grid system or census tracts. We consider M heterogeneous data modalities that describe these regions, such as Points-of-Interest (POI), land use, and human mobility. The set of modality indices is denoted as $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$.

Input. For each modality $v \in \mathcal{V}$, its corresponding feature matrix is denoted as $\mathbf{X}^v \in \mathbb{R}^{N \times d_v}$, where each row $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$ represents the feature vector of region r_i under modality v. The multi-view features for region r_i are therefore represented as $\mathbf{X}_i = \{\, \mathbf{x}_i^v \mid v \in \mathcal{V} \,\}$.

Downstream Tasks. We consider a diverse set of S downstream urban tasks, $\mathcal{T} = \{T_1, T_2, \ldots, T_S\}$. Each task T_s is associated with a ground-truth label y_i^s for each region r_i. The label can be a continuous value for regression tasks (e.g., crime rate, $y_i^s \in \mathbb{R}$) or a discrete value for classification tasks (e.g., land use type, $y_i^s \in 0, 1, \ldots, C-1$).

Problem Formulation. Given the multi-view features $\mathbf{X}_i$ of a set of N regions and a downstream task T_s, the goal of urban region representation learning is to learn an encoder f that maps each region to a d-dimensional embedding $\mathbf{e}_i = f(\mathbf{X}_i)$. The embedding matrix is denoted by $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N] \in \mathbb{R}^{N \times d}$. A lightweight predictor g then outputs the estimated label $\hat{y}_i^s = g(\mathbf{e}_i, T_s)$. The overall objective is to minimize the discrepancy between the predicted and ground-truth labels across all tasks.

3.2 Interpretable Universal Representation Learning

Interpretable universal representation learning is dedicated to learning a universal region representation that is both powerful and inherently interpretable. The central idea is to explicitly disentangle the information contributions from different data views, a concept inspired by PID. We operationalize this idea through a structured two-step process. First, the multi-view feature learning step aims to generate view-specific embeddings from raw data. These embeddings then serve as the input to our core innovation, the MoFE module, which is designed to fuse and disentangle them into the universal interpretable representation.

3.2.1 Multi-view Feature Learning. We first generate view-specific representations from the raw multimodal urban data. Following [15], we adopt a Transformer-based encoder for each data view. Given the input feature matrices for POI, land use, and mobility, which are denoted as $\mathbf{X}^p, \mathbf{X}^l, \mathbf{X}^m$ respectively, the encoder leverages self-attention mechanisms to capture intra-view spatial correlations. This process yields a set of information-rich, view-specific embedding matrices $\{\mathbf{Z}^p, \mathbf{Z}^l, \mathbf{Z}^m\}$, which serve as the foundational input for the subsequent fusion stage.

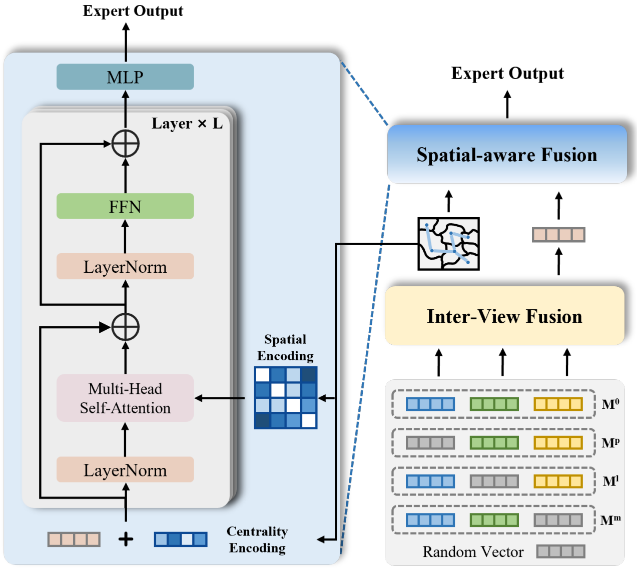

Figure 2: Internal architecture of a fusion expert.

3.2.2 Mixture of Fusion Experts (MoFE). To fuse the multi-view embeddings in an interpretable manner, we introduce the MoFE module. MoFE comprises a set of specialized experts and an adaptive router that assigns weights to them. We first describe the sophisticated internal architecture shared by all experts.

Internal Architecture of a Fusion Expert. Each expert in our framework, regardless of its specialization, is a powerful neural network designed to process multi-view features and capture spatial dependencies. As detailed in Figure 2, this architecture consists of two sequential components:

- **Inter-View Fusion:** To combine information from different modalities, we adopt the ViewFusion mechanism [15]. Given a set of input view embeddings $\{\mathbf{Z}^v\}$, this module learns adaptive fusion weights α_v for each view v by computing pairwise correlations. The output is a single fused representation $\mathbf{Z} = \sum_v \alpha_v \cdot \mathbf{Z}^v$.
- **Spatial-aware Fusion:** To explicitly model spatial autocorrelation, the fused representation $\mathbf{Z}$ is then processed by a spatial-aware fusion module. This module enhances $\mathbf{Z}$ with structural information derived from a spatial adjacency matrix $\mathbf{A}$ and applies a multi-head self-attention mechanism with a spatial bias. The attention score between regions i and j is formulated as:

$$\alpha_{ij} = \text{Softmax}_j\left(\frac{(\mathbf{W}_Q\hat{\mathbf{Z}}_i)(\mathbf{W}_K\hat{\mathbf{Z}}_j)^\top}{\sqrt{d}} + \beta B_{ij}\right), \qquad (1)$$

where $\hat{\mathbf{Z}}_i$ and $\hat{\mathbf{Z}}j$ are the i-th and j-th row vectors of the spatially-enhanced feature matrix $\hat{\mathbf{Z}}$, respectively, β is the hyperparameter, and B_{ij} is a spatial bias term from $\mathbf{A}$. Stacking L such layers yields the final expert output, which we denote as $\tilde{\mathbf{E}}_k$ for an expert k.

Expert Composition and Adaptive Routing. Our MoFE implementation utilizes five experts built upon this architecture: three **Uniqueness Experts** (for POI, Land-use, Mobility), one **Synergy Expert**, and one **Redundancy Expert**. Given the view-specific embeddings $\{\mathbf{Z}^p, \mathbf{Z}^l, \mathbf{Z}^m\}$, we obtain an "anchor" output $\tilde{\mathbf{E}}_k$ from each of the five experts k. An MLP-based router then computes an importance weight w_k for each expert based on the input features. The final universal region embedding $\mathbf{E}$ is the weighted aggregation of all expert outputs:

$$\mathbf{E} = \sum_{k\in\{p,l,m,syn,red\}} w_k \cdot \tilde{\mathbf{E}}_k. \qquad (2)$$

Training Strategy for Expert Specialization. To ensure each expert plays its intended role, we employ a specialized training strategy rooted in the Triplet Margin Loss [13] framework. The core idea is to adaptively define positive and negative samples for each expert based on its target interaction type. Let $\tilde{\mathbf{E}}^0_k$ be the anchor output from expert k using the full input, and $\tilde{\mathbf{E}}^v_k$ be the output when view v is masked. We employ the Cosine Distance as our distance metric $d(\cdot,\cdot)$. The specific loss functions are formulated as follows:

- **Uniqueness Loss ($\mathcal{L}_{uni}$):** For expert $uni, uni \in \{p, l, m\}$, we use the classic triplet loss, where $\tilde{\mathbf{E}}^{uni}_{uni}$ is the negative sample and all other masked outputs $\tilde{\mathbf{E}}^v_{uni}$ $(v \neq uni)$ are positive samples.

$$\mathcal{L}_{uni} = \sum_{v\neq uni} \max\left(d(\tilde{\mathbf{E}}^0_{uni}, \tilde{\mathbf{E}}^v_{uni}) - d(\tilde{\mathbf{E}}^0_{uni}, \tilde{\mathbf{E}}^{uni}_{uni}) + \gamma, 0\right), \qquad (3)$$

where γ is the margin of triplet loss. This loss enforces that removing the expert's own view leads to the largest embedding change, while removing other views leads to minimal changes.

- **Synergy Loss ($\mathcal{L}_{syn}$):** For the synergy expert syn, all masked outputs are negatives. This corresponds to a triplet loss variant with no external positive samples.

$$\mathcal{L}_{syn} = \sum_v \max\left(1 - d(\tilde{\mathbf{E}}^0_{syn}, \tilde{\mathbf{E}}^v_{syn}), 0\right). \qquad (4)$$

This loss aims to maximize the distance $d(\tilde{\mathbf{E}}^0_{syn}, \tilde{\mathbf{E}}^v_{syn})$, thereby encouraging the model to be highly sensitive to the removal of any single view.

- **Redundancy Loss ($\mathcal{L}_{red}$):** For the redundancy expert red, all masked outputs are positives. This degenerates the triplet framework into a purely attractive loss with no negative samples.

$$\mathcal{L}_{red} = \sum_v d(\tilde{\mathbf{E}}^0_{red}, \tilde{\mathbf{E}}^v_{red}). \qquad (5)$$

The total specialization loss is $\mathcal{L}_{\text{MoFE}} = \sum_{uni\in\{p,l,m\}} \mathcal{L}_{uni} + \mathcal{L}_{syn} + \mathcal{L}_{red}$. This principled training strategy ensures the interpretability of the learned expert weights.

3.3 Prompt4RE

While the MoFE module yields a powerful and interpretable universal region embedding $\mathbf{E}$, this representation is inherently task-agnostic. To unlock its full potential for diverse downstream applications, a task-specific adaptation mechanism is crucial. To this end, Prompt4RE is designed to adapt the universal embedding $\mathbf{E}$ in a semantically guided manner. Prompt4RE leverages the advanced reasoning capabilities of a frozen MLLM to bridge the gap between high-level task descriptions and low-level representation tuning.

Semantic Prompt Initialization via MLLM.. To overcome the limitations of task-agnostic prompt initialization, we leverage a frozen MLLM, specifically `LLaMA-3.2-11B-Vision-Instruct` [5], as a zero-shot semantic engine. For each downstream task (e.g., crime prediction), we provide the MLLM with a carefully designed prompt template (see Appendix A for an example) and relevant multimodal urban data, including satellite images, street-view photos, and geo-textual descriptions (Appendix B). The MLLM processes these inputs and generates a rich, task-relevant analysis. We then extract the MLLM's final hidden state, which encapsulates this contextual knowledge, and use it as the initial prompt embedding, denoted as **P**. This method initializes the prompt embedding, **P**, with task-specific semantics, ensuring it is aligned with the requirements of the downstream task from the outset.

Prompt-Representation Alignment. With a semantically rich prompt embedding **P** and the universal region embedding **E** from MoFE, the next step is to align them. To this end, we propose the P-R Alignment module. The core idea is to let the region embedding "query" the task-specific semantic knowledge contained within the prompt. This is implemented using a multi-head cross-attention mechanism, where the queries (**Q**) are derived from the universal region embedding **E**, while the keys (**K**) and values (**V**) are derived from the prompt embedding **P**. For a single attention head i, the computation is as follows:

$$\mathrm{att}_i = \mathrm{Softmax}\left(\frac{(\mathbf{E}\mathbf{W}_Q^{(i)})(\mathbf{P}\mathbf{W}_K^{(i)})^\top}{\sqrt{d_i}}\right)\mathbf{P}\mathbf{W}_V^{(i)}, \tag{6}$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)}$ are learnable projection matrices for head i, and d_i is the head's dimension. The outputs of all H heads are then concatenated:

$$\mathrm{MHA}(\mathbf{E}, \mathbf{P}) = \mathrm{Concat}(\mathrm{att}_1, \dots, \mathrm{att}_H). \tag{7}$$

To preserve the original spatial semantics learned in the first stage, we incorporate a residual connection and layer normalization, producing the final aligned prompt embedding $\mathbf{P}'$:

$$\mathbf{P}' = \mathrm{LayerNorm}(\mathrm{MHA}(\mathbf{E}, \mathbf{P}) + \mathbf{E}\mathbf{W}_{\mathrm{res}}), \tag{8}$$

where $\mathbf{W}_{\mathrm{res}}$ is a learnable projection for the residual path.

Task-Specific Prediction and Optimization. The aligned prompt $\mathbf{P}'$ is then passed through a lightweight fully connected layer to transform it into a soft prompt **S** with the same dimension as the region embedding **E**. The final task-specific representation **H** is formed by concatenating the universal embedding and the soft prompt: $\mathbf{H} = \mathbf{E} \parallel \mathbf{S}$, where $\parallel$ denotes concatenation. For each region r_i, its enhanced representation $h_i \in \mathbf{H}$ is fed into a simple Feed-Forward Network (FNN) to produce the final prediction $\hat{y}_i = \mathrm{FNN}(h_i)$. The Prompt4RE module is trained end-to-end by minimizing the Mean Squared Error (MSE) loss between the predictions and the ground-truth labels:

$$\mathcal{L}_{\mathrm{Prompt4RE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{9}$$

It is important to note that during this second stage of training, the parameters of the MoFE module and the view-specific encoders are frozen, ensuring that only the lightweight prompt-related components are tuned.

3.4 Model Training and Optimization

The training of RegionMoFE follows a two-stage paradigm: first learning a universal representation, then adapting it for downstream tasks.

Stage 1: Interpretable Universal Representation Learning. The primary goal of the first stage is to train the view-specific encoders and the MoFE to produce a universal region embedding **E** that is both powerful and interpretable. To achieve this, we optimize a composite loss function, $\mathcal{L}_{\mathrm{Stage1}}$, which consists of two main components: a multi-task reconstruction loss $\mathcal{L}_{\mathrm{task}}$ to ensure representation quality, and our previously defined expert specialization loss $\mathcal{L}_{\mathrm{MoFE}}$ to enforce interpretability.

The multi-task loss $\mathcal{L}_{\mathrm{task}}$ comprises three self-supervised objectives, each designed to ensure the embedding **E** captures essential information from a specific data view:

- **POI/Land-use Similarity Reconstruction:** To preserve semantic similarity, we decode the universal embedding **E** back into view-specific embeddings ($\mathbf{E}^p, \mathbf{E}^l$) using view-specific MLPs. We then enforce that the inner product of these decoded embeddings reconstructs the pre-computed region similarity matrices ($\mathbf{A}^p, \mathbf{A}^l$). The loss for view $v \in \{p, l\}$ is:

$$\mathcal{L}_v = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|\mathbf{A}_{i,j}^v - \mathbf{e}_i^v \cdot (\mathbf{e}_j^v)^\top\right\|_2^2. \tag{10}$$

- **Mobility Distribution Reconstruction:** To capture mobility patterns, we aim to reconstruct the empirical transition probability distributions between regions. The universal embedding **E** is projected into a source space $\mathbf{E}^S$ and a destination space $\mathbf{E}^D$. The loss $\mathcal{L}_m$ is defined as the KL-divergence between the predicted and true distributions for both outflow and inflow traffic, encouraging the model to understand regional connectivity.

The total multi-task loss is $\mathcal{L}_{\mathrm{task}} = \mathcal{L}_p + \mathcal{L}_l + \mathcal{L}_m$. The final loss for Stage 1 dynamically combines this with the expert loss:

$$\mathcal{L}_{\mathrm{Stage1}} = \lambda\mathcal{L}_{\mathrm{task}} + (1-\lambda)\mathcal{L}_{\mathrm{MoFE}}, \tag{11}$$

where λ is a learnable parameter that automatically balances the trade-off between representation quality and expert specialization.

Stage 2: Prompt4RE. In the second stage, we focus on adapting the learned universal embedding **E** for a specific downstream task. During this stage, all parameters of the view-specific encoders and the MoFE module are frozen. This ensures that the rich, general knowledge captured in Stage 1 is preserved, and makes the adaptation process highly efficient and lightweight.

The training objective is to optimize only the parameters of the Prompt4RE module. This is achieved by minimizing the task-specific loss $\mathcal{L}_{\mathrm{Prompt4RE}}$, which measures the discrepancy between the model's final predictions and the ground-truth labels for the target task. This focused optimization allows the model to learn a task-specific soft prompt **S** that effectively queries and refines the universal embedding for optimal performance on the task at hand.

4 EXPERIMENTS

In this section, we conduct extensive experiments to address the following research questions: (1) **RQ1**: Can RegionMoFE outperform prior methods and generalize well across diverse urban computing tasks? (2) **RQ2**: How does each component contribute to RegionMoFE? (3) **RQ3**: Does Prompt4RE have generalization ability? (4) **RQ4**: What is the importance of the information captured by each expert across different tasks? (5) **RQ5**: How do key hyperparameters affect the model's performance?

4.1 Experimental Setup

4.1.1 Datasets. We use real-world data from New York City [11] and Chicago [1]. Following [15], we adopt POIs, land use, and human mobility data for the first-stage training. To support the Prompt4RE module, we further construct multimodal inputs for each region: (1) satellite and street-view images retrieved from Google Maps [4]; and (2) geospatial descriptions, including latitude/longitude, addresses, and POIs, collected from OpenStreetMap [12] and aggregated into textual geographic descriptions. For evaluation, we follow [14] and use four benchmark downstream tasks conducted on these regions. Detailed dataset statistics are reported in Appendix C.

4.1.2 Baselines. We compare RegionMoFE with a broad range of state-of-the-art baselines, categorized into one-stage and two-stage methods. For all baselines, we adopt the hyperparameter settings reported in their original papers.

One-stage Methods: (1) **MVURE** [30]: Employs a Graph Attention Network (GAT) on multi-view graphs (mobility, POIs) with an adaptive fusion module. (2) **MGFN** [23]: Focuses on constructing and fusing multiple temporal mobility graphs to capture spatiotemporal patterns. (3) **ReCP** [9]: A multi-view framework based on information consistency, using cross-view contrastive learning. (4) **HAFusion** [15]: Proposes a dual-attention fusion module to capture higher-order associations across both views and regions.

Two-stage Methods: (1) **HREP** [32]: Learns general embeddings from a heterogeneous graph and then applies a prefix-tuning strategy with continuous prompts for task adaptation. (2) **UrbanCLIP** [27]: Generates textual descriptions for satellite images via a MLLM and aligns visual features with the generated text using contrastive pre-training. (3) **FlexiReg** [14]: Learns from fine-grained grid cells and uses a "Prompt Enhancer" that incorporates features from a large language model and street-view images for task-specific customization.

4.1.3 Implementation Details and Metrics. Our model is trained in two stages, each for 3000 epochs. For interpretable universal representation learning , we set the learning rate to 5e-4/4e-4 and the triplet loss margin to 0.9/0.8 for NYC/Chicago, respectively. The Spatial-aware Fusion module uses 3 layers with 4 heads ($\beta = 0.5$) for NYC, and 4 layers with 11 heads ($\beta = 0.3$) for Chicago. For Prompt4RE, the soft prompt dimension matches the embedding dimension (144). The learning rate is set to 1e-4 for crime and service call tasks, and 3e-4 for check-in and population tasks across both cities. All hyperparameters were determined by grid search on a validation set. For performance evaluation, we adopt three standard metrics consistent with prior work [14, 15, 32]: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

4.2 Performance Comparison (RQ1)

We compare RegionMoFE with state-of-the-art baselines across four downstream tasks in two cities. The comprehensive results are presented in Table 1. As confirmed by t-tests ($p < 0.05$ against the strongest baseline on all tasks), RegionMoFE demonstrates statistically significant improvements, establishing a new state of the art. Our key findings are as follows:

(1) RegionMoFE consistently and substantially outperforms all baselines. Across all 24 metrics (2 cities × 4 tasks × 3 metrics), RegionMoFE achieves the best performance. The average improvement over the strongest baseline is significant, reaching 8.5% in New York City and a more pronounced 15.4% in Chicago. This superior performance underscores the power of our RegionMoFE.

(2) The advantage of RegionMoFE is particularly prominent in semantically rich tasks. We observe the most gains in tasks like Check-in and Population prediction. For instance, in the Chicago Check-in task, RegionMoFE slashes the RMSE by a remarkable 45.3% compared to the best baseline. We attribute this to our Prompt4RE module, where MLLM-generated semantic prompts help the model comprehend the task's essence, thereby aligning the universal representation more effectively with its specific requirements.

(3) RegionMoFE achieves superior performance compared to all baselines, most notably outperforming FlexiReg, the strongest competitor which also incorporates a LLM. This performance advantage is quantitatively evident across all tasks; for instance, on the NYC Population task, RegionMoFE improves the R^2 score from 0.701 (FlexiReg) to 0.745. We attribute this to our Prompt4RE module, which leverages MLLM's world knowledge to generate task-specific semantic prompts. These prompts help the model better understand the underlying urban functions (e.g., commercial vs. residential), an area where methods relying solely on structural data struggle.

(4) The two-stage method demonstrates clear advantages. Both RegionMoFE and FlexiReg generally outperform one-stage methods like HAFusion and MVURE. This confirms the benefit of decoupling general representation learning from task-specific adaptation. However, our model pushes this paradigm further by introducing a more sophisticated fusion backbone and a more direct prompting mechanism, leading to its superior performance.

4.3 Ablation Studies (RQ2)

We conduct ablation studies to assess the contributions of RegionMoFE's key components. We design three variants by removing or replacing core modules: w/o MoFE: To evaluate the core Mixture-of-Fusion-Experts design, we replace the entire MoFE module with a single, unified fusion block. w/o SpatialFusion: We remove the Spatial-aware Fusion module to isolate the impact of modeling spatial autocorrelation through geographical neighbors. w/o Prompt4RE: We remove the lightweight Prompt4RE module to assess the value of task-specific adaptation. The model in this case relies solely on the representation from stage 1.

Table 1: Performance comparison on four downstream tasks. Best: Bold, Second: Underline. * indicates a statistically significant improvement with p-value < 0.05. The 'Improvement' is calculated against the best baseline.

New York City	Crime			Check-in			Service Call			Population		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	67.4	90.5	0.625	285.1	461.0	0.682	1402	2128	0.398	2899	3708	0.508
MGFN	73.2	91.4	0.618	345.0	503.5	0.621	1653	2250	0.327	3222	4207	0.367
ReCP	81.4	101.3	0.483	233.8	392.7	0.763	1478	2136	0.366	3527	4434	0.329
HAFusion	56.1	76.1	0.734	202.8	322.8	0.844	1273	1951	0.493	2497	3277	0.616
HREP	66.1	84.4	0.674	274.1	417.7	0.739	1396	2011	0.462	3118	4023	0.421
UrbanCLIP	97.4	126.1	0.267	393.6	602.4	0.458	1409	2401	0.232	3338	4499	0.276
FlexiReg	53.2	73.6	0.751	198.2	309.0	0.850	1303	1989	0.497	2231	2974	0.701
RegionMoFE*	**52.1**	**69.0**	**0.776**	**185.4**	**251.3**	**0.905**	**1201**	**1732**	**0.574**	**2053**	**2627**	**0.745**
Improvement	**2.1%**	**6.2%**	**3.3%**	**6.5%**	**18.7%**	**6.5%**	**5.6%**	**11.2%**	**15.5%**	**8.0%**	**11.7%**	**6.3%**

Chicago	Crime			Check-in			Service Call			Population		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	100.4	129.2	0.461	1693	3171	0.656	190.3	266.9	0.441	13717	17174	0.313
MGFN	107.4	137.9	0.386	1281	2276	0.817	208.2	293.4	0.329	13071	16578	0.359
ReCP	86.9	120.1	0.534	1272	2341	0.804	206.7	303.4	0.284	12085	17029	0.325
HAFusion	77.8	107.1	0.631	929	1947	0.870	159.3	222.0	0.613	10678	13988	0.544
HREP	88.3	114.4	0.578	1679	3135	0.664	185.7	262.2	0.468	12063	15397	0.447
UrbanCLIP	101.6	134.7	0.416	2612	4885	0.186	183.2	256.3	0.491	13328	17498	0.288
FlexiReg	61.7	85.1	0.766	922	1775	0.891	121.1	178.2	0.753	8126	11395	0.698
RegionMoFE*	**59.5**	**78.4**	**0.807**	**607**	**971**	**0.968**	**106.4**	**164.6**	**0.781**	**6552**	**9034**	**0.815**
Improvement	**3.6%**	**7.9%**	**5.4%**	**34.1%**	**45.3%**	**8.6%**	**12.1%**	**7.6%**	**3.7%**	**19.4%**	**20.7%**	**16.8%**

Table 2: Ablation Study.

NYC	Crime			Service Call		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
RegionMoFE	**52.1**	**69.0**	**0.776**	**1201**	**1732**	**0.574**
w/o Prompt4RE	55.5	73.0	0.756	1322	1878	0.531
w/o Spatial Fusion	58.1	76.2	0.735	1249	1887	0.526
w/o MoFE	59.0	79.1	0.714	1262	1955	0.492

The results, summarized in Table 2, lead to several key insights. First, removing the MoFE module (w/o MoFE) causes the most significant performance degradation across all metrics (e.g., Crime MAE increases from 52.1 to 59.0). This confirms that explicitly modeling data heterogeneity via specialized experts is crucial for effective fusion, far superior to a monolithic approach. Second, the performance drop in the w/o SpatialFusion variant demonstrates that injecting geographical context is essential for spatially dependent tasks like Crime and Service Call prediction. Finally, the w/o Prompt4RE variant offers a dual insight. While the performance drop upon its removal confirms the value of task-specific adaptation from Prompt4RE, it is telling that this variant remains the strongest among all ablations. This demonstrates that Stage 1 alone produces a high-quality universal representation, which Prompt4RE then effectively refines.

4.4 Generalizability of Prompt4RE (RQ3)

To demonstrate that Prompt4RE is a model-agnostic, plug-and-play module, we integrated it with various baseline methods by using their final region embeddings as input to our prompting mechanism. For models with existing prompt modules like HREP and FlexiReg, we performed a direct replacement. As shown in Table 3, the results are compelling: Prompt4RE consistently boosts the performance of all baseline models, achieving average improvements of 6.4% on Crime, 9.3% on Check-in, and 9.5% on Service Call tasks. This experiment validates that Prompt4RE can function as a universal semantic enhancement layer, effectively injecting MLLM-derived world knowledge into diverse urban representation learning frameworks, regardless of their underlying architecture.

4.5 Interpretability Analysis (RQ4)

To further understand how RegionMoFE utilizes different data modalities, we conduct an interpretability analysis based on the expert weights learned by the MoFE (Figure 3). Several clear patterns emerge. (1) Dominant signals for simple tasks. Population and Check-in predictions rely mainly on the Mobility and POI experts, matching intuitive domain knowledge that population flow and check-ins reflect mobility and POI density. (2) Cross-modal composition for complex tasks. For Crime and Service Call, multiple experts—POI, Land-use, and Mobility—receive comparable

Table 3: Performance boost when equipping existing models with our Prompt4RE module. Models with ‘P’ denote the original baseline enhanced by Prompt4RE. ‘Average Improvement’ is the mean gain over all baselines.

NYC	Crime			Check-in			Service Call		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
MVURE	67.4	90.5	0.625	285.1	461.0	0.682	1402	2128	0.398
MVURE$_P$	63.4	85.2	0.668	264.6	422.2	0.734	1269	1963	0.488
MGFN	73.2	91.4	0.618	345.0	503.5	0.621	1653	2250	0.327
MGFN$_P$	67.5	85.4	0.666	285.2	434.1	0.718	1530	2146	0.387
ReCP	81.4	101.3	0.483	233.8	392.7	0.763	1478	2136	0.366
ReCP$_P$	74.6	95.3	0.522	208.9	331.4	0.811	1382	1917	0.455
HAFusion	56.1	76.1	0.734	202.8	322.8	0.844	1273	1951	0.493
HAFusion$_P$	52.7	69.7	0.770	187.0	277.3	0.885	1191	1888	0.526
HREP	66.1	84.4	0.674	274.1	417.7	0.739	1396	2011	0.462
HREP$_P$	59.5	79.1	0.714	251.4	360.3	0.806	1247	1909	0.516
FlexiReg	53.2	73.6	0.751	198.2	309.0	0.850	1303	1989	0.497
FlexiReg$_P$	51.5	70.3	0.769	194.8	298.237	0.867	1220	1903	0.522
Average Improvement	**6.9%**	**6.2%**	**6.0%**	**8.8%**	**11.5%**	**7.6%**	**7.8%**	**5.9%**	**14.8%**

weights, showing that RegionMoFE captures synergistic rather than single-source interactions. (3) Adaptive pruning of irrelevant experts. The routing gate down-weights unhelpful components such as the Synergy expert in simple tasks, highlighting selective focus.

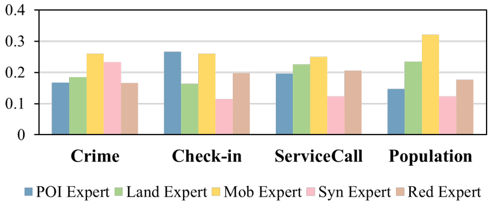

Figure 3: Learned expert weights for each downstream task. The y-axis indicates the assigned weight for each expert.

4.6 Hyperparameter Sensitivity Analysis (RQ5)

We investigate the impact of two key hyperparameters on the Chicago dataset, as shown in Figure 4. For the Prompt4RE module, we examine how the dimensionality of the soft prompt affects prompt learning performance. We observe that our model remains relatively stable across different dimensions, with the 144-dimensional soft prompt achieving slightly better results than others. For the MoFE module, we study the effect of varying the number of layers in the spatial fusion module during the first training stage. The results indicate that choosing an appropriate number of spatial fusion layers is critical. In both crime prediction and population prediction tasks, performance drops significantly when the number of layers is reduced to 2, suggesting that too few layers fail to capture sufficient spatial knowledge. Conversely, when the number of layers exceeds 4, performance also degrades, likely due to overfitting introduced by excessive spatial fusion layers.

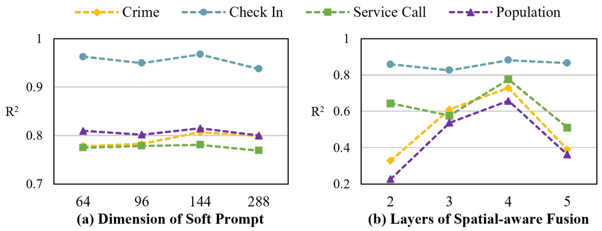

Figure 4: Hyperparameter sensitivity analysis.

5 CONCLUSION AND FUTURE WORK

In this paper, we presented RegionMoFE for urban region representation learning. RegionMoFE unifies interpretable fusion and semantically guided adaptation, addressing two long-standing limitations: the black-box nature of fusion and the task-agnosticity of adaptation. Through the MoFE, our first-stage interpretable universal representation learning explicitly disentangles heterogeneous interactions into unique, redundant, and synergistic components, providing transparent and quantifiable interpretability. Building upon this foundation, the Prompt4RE module leverages a frozen MLLM to produce semantically grounded prompts that align representation learning directly with downstream task intents. Extensive experiments across four diverse urban tasks demonstrate that RegionMoFE consistently outperforms current state-of-the-art methods, while also providing interpretable insights into how different data modalities contribute to task performance. Beyond improved accuracy, RegionMoFE establishes a general paradigm for bridging interpretable information-theoretic modeling and language-guided adaptation in multi-view learning. In future work, we plan to extend RegionMoFE to enable spatiotemporal reasoning for urban forecasting. Furthermore, we will explore cross-city generalization through zero-shot semantic adaptation across heterogeneous regions.

References

[1] Chicago Government. 2025. *Chicago Data Portal*. https://data.cityofchicago.org/ Retrieved June 28, 2025 from https://data.cityofchicago.org/.

[2] Foursquare. 2025. *Foursquare*. https://foursquare.com/ Retrieved June 28, 2025 from https://foursquare.com/.

[3] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Xiaolin Li. 2019. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 906–913.

[4] Google Maps. 2025. *Google Maps*. https://www.google.com/maps Retrieved June 28, 2025 from https://www.google.com/maps.

[5] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[6] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2025. Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28061–28069.

[7] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised representation learning of spatial data via multimodal embedding. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1993–2002.

[8] Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. Urban region representation learning with openstreetmap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1363–1373.

[9] Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. 2024. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8724–8732.

[10] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems* 36 (2023), 27351–27393.

[11] NYC Government. 2025. *NYC Open Data*. https://opendata.cityofnewyork.us/ Retrieved June 28, 2025 from https://opendata.cityofnewyork.us/.

[12] OpenStreetMap. 2025. *OpenStreetMap*. https://www.openstreetmap.org/ Retrieved June 28, 2025 from https://www.openstreetmap.org/.

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[14] Fengze Sun, Yanchuan Chang, Egemen Tanin, Shanika Karunasekera, and Jianzhong Qi. 2025. FlexiReg: Flexible Urban Region Representation Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2702–2713.

[15] Fengze Sun, Jianzhong Qi, Yanchuan Chang, Xiaoliang Fan, Shanika Karunasekera, and Egemen Tanin. 2024. Urban region representation learning with attentive fusion. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 4409–4421.

[16] Tycho MS Tax, Pedro AM Mediano, and Murray Shanahan. 2017. The partial information decomposition of generative neural network models. *Entropy* 19, 9 (2017), 474.

[17] Hongjian Wang and Zhenhui Li. 2017. Region representation learning via mobility flow. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 237–246.

[18] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1013–1020.

[19] Charles Westphal, Stephen Hailes, and Mirco Musolesi. 2024. Partial Information Decomposition for Data Interpretability and Feature Selection. *arXiv preprint arXiv:2405.19212* (2024).

[20] Michael Wibral, Viola Priesemann, Jim W Kay, Joseph T Lizier, and William A Phillips. 2017. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain and cognition* 112 (2017), 25–38.

[21] Paul L Williams and Randall D Beer. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* (2010).

[22] WorldPop. 2025. *WorldPop*. https://www.worldpop.org/ Retrieved June 28, 2025 from https://www.worldpop.org/.

[23] Shangbin Wu, Xu Yan, Xiaoliang Fan, Shirui Pan, Shichao Zhu, Chuanpan Zheng, Ming Cheng, and Cheng Wang. 2022. Multi-graph fusion networks for urban region embedding. *arXiv preprint arXiv:2201.09760* (2022).

[24] Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. 2024. Refound: Crafting a foundation model for urban region understanding upon language and visual foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3527–3538.

[25] Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. 2025. I2MoE: Interpretable Multimodal Interaction-aware Mixture-of-Experts. *arXiv preprint arXiv:2505.19190* (2025).

[26] Zhuo Xu and Xiao Zhou. 2024. CGAP: urban region representation learning with coarsened graph attention pooling. *arXiv preprint arXiv:2407.02074* (2024).

[27] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*. 4006–4017.

[28] Zijun Yao, Yanjie Fu, Bin Liu, Wangsu Hu, and Hui Xiong. 2018. Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

[29] Liang Zhang, Cheng Long, and Gao Cong. 2022. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9031–9036.

[30] Mingyang Zhang, Tong Li, Yong Li, and Pan Hui. 2021. Multi-view joint graph representation learning for urban region embedding. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 4431–4437.

[31] Yunchao Zhang, Yanjie Fu, Pengyang Wang, Xiaolin Li, and Yu Zheng. 2019. Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1700–1708.

[32] Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. 2023. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 4981–4989.

A Task Prompt Template

We provide an illustrative example of the prompt (Figure 5) used in the crime prediction task within Prompt4RE. The prompt is structured into four components: (i) role specification, (ii) input information description, (iii) task definition, and (iv) reasoning guidance with output format. This design encourages the large language model to perform step-by-step reasoning and to produce rich, structured semantic outputs.

B Geographic Information Description

We also present an example of the geographic information description, show as Figure 6. This template follows the setting of [14] and includes latitude–longitude coordinates, address information, and categories of Points of Interest (POIs) contained within the target region.

C Dataset Statistics

Table 4: Dataset Statistics.

Dataset	NYC	CHI	Source
Regions	180	77	Open Portal [1, 11]
POIs	24496	57891	OpenStreetMap [12]
POI categories	26	26	OpenStreetMap [12]
Land-use categories	11	12	OpenStreetMap [12]
Taxi trips	10953879	3381807	Open Portal [1, 11]
Satellite images	180	77	Google Maps [4]
Street view photos	11629	5005	Google Maps [4]
Crime records	35335	18200	Open Portal [1, 11]
Check-in counts	106902	167232	Foursquare [2]
Service call records	516187	24350	Open Portal [1, 11]
Population counts	1540692	2508984	WorldPop [22]

You are a top expert in urban criminology and urban planning. Your task is to analyze various data of a city area and deeply understand the potential characteristics related to crime risk in that area.

[Input Information]
Given are three types of information:
1. Satellite images: Please focus on the macroscopic layout of the area, such as: land use types (residential areas, commercial areas, industrial areas, open spaces), building density, road network structure (grid-like or disordered), connectivity with surrounding areas (such as central commercial areas, transportation hubs), green coverage rate, etc.
2. Street view images: Please focus on the micro environment of the area, such as: the condition of buildings (such as graffiti on walls, broken windows), the condition of public facilities (such as whether the streetlights are intact, whether there are surveillance cameras), the cleanliness of the streets, accessibility of sidewalks, signs of commercial activities (such as busy shops or closed and bankrupt ones), etc.
3. Text information: Contains the latitude and longitude of the area, the address, and POI category information.

[Your Core Task]
Please conduct a step-by-step reasoning to analyze which features and patterns in the multi-modal information are highly correlated with potential urban crime risk. Combine all the information to conduct a comprehensive 'region diagnosis'.

[Please think and output in the following structure]
1. Physical environment analysis: Based on satellite and street view images, list all the physical features that may indicate an increase or decrease in crime risk (such as： blind spots without natural surveillance, abandoned buildings, poorly maintained public spaces).
2. Social and economic background analysis: Based on text data and the inferred information in the images (such as the level of regional prosperity), analyze the possible social and economic factors that may affect the crime rate (such as: signs of economic recession, population mobility, community cohesion).
3. Risk comprehensive assessment: Combine all these points to make a preliminary judgment of the overall crime risk level (high, medium, low), and explain your core reasons. Focus on how the different modalities of information mutually confirm or provide new perspectives.
4. Conclusion and feature extraction: Finally, please summarize a set of key feature labels that best represent the crime prediction risk of the area (such as: #High-density old residences #Insufficient commercial vitality #Insufficient lighting #Low-income population concentration #Near transportation hubs).

Please focus your thinking on the structured analysis above. Your final output should be this detailed analysis report.

Figure 5: Task prompt in crime prediction task.

The following is the geographical information description within the area.

Centroid Coordinates: (40.75194855683927,-73.98411036929014).

Address: 29, West 38th Street, Midtown South, Manhattan Community Board 5, Manhattan, New York County, City of New York, New York, 10018, United States.

Included POIs: 3 educational institutions (1.79%), 9 commercial and industrial properties (5.36%), 19 accommodation (11.31%), 2 cultural and recreational venues (1.19%), 3healthcare and medical facilities (1.79%), 10 entertainment venues (5.95%), 1 places of worship (0.6%), 60 food and drink establishments (35.71%), 12 parking facilities (7.14%),1 transportation and transit facilities (0.6%), 3 financial ser-vices (1.79%), 45 others. The total number of POIs in this region is 168.

Figure 6: An example of geographic information description.

The statistics and sources of the datasets are summarized in Table 4. Region partitioning is based on census tracts for New York City and community boundaries for Chicago. The satellite images are of size 640×640 pixels, while the street-view images are of size 2048×1024 pixels. For New York City, which contains 180 regions, we uniformly sample 65 street-view images per region, constrained by the input limitations of large language models. For three smaller regions, where fewer images are available, we instead use the maximum available numbers: 60, 34, and 30 images, respectively. For Chicago, which consists of 77 regions, we uniformly sample 65 street-view images per region.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009